**Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/**
**Journal of Xi'an Shiyou University, Natural Sciences Edition**
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 02 | 2025
DOI: 10.5281/zenodo.14842473

# AN ENHANCED SYSTEM FOR PREDICTING HEART STROKE USING MACHINE LEARNING

**S MUTHUKUMAR**
Assistant Professor, Department of Computer Science Engineering, St. Joseph’s Institute of Technology, Chennai, India. Email: smk252it2021@gmail.com

**Dr. KALAI VANI YS**
Assistant Professor, Department of Information Science and Engineering, BMS Institute of Science and Technology & Management, Bangalore, India. Email: kalaivaniys@bmsit.in

**HEMAL BABU H**
Assistant Professor, Department of Artificial Intelligence & Data Science, Rajalakshmi Institute of Technology, Chennai, India. Email: hemalbabuh2k11@gmail.com

**MALATHI S**
Professor, Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, India. Email: malathi.raghuram@gmail.com

**Dr. D. MENAGA**
Associate Professor, Department of Computer Science Engineering, St. Joseph’s Institute of Technology, Chennai, India. Email: dev.menaga@gmail.com

**S ABHIRAMI**
Assistant Professor, Department of Mathematics, Sona College of Technology, Salem, India.
Email: abhiramishanmugasundaram@gmail.com

### Abstract

Globally, heart disease and strokes are now among the top causes of death, particularly for adults. According to the WHO's data analysis, 17.9 million persons worldwide are estimated to have died from cardiovascular disease (CVD) in 2019. Heart attacks and strokes were responsible for almost 85% of all deaths. Hence, it is essential to identify cardiovascular diseases (CVD) so that the right treatments can be given as soon as feasible. Age, gender, hypertension, heart disease, type of work, type of residence, average blood glucose level, BMI, and smoking status were the parameters we collected from the Kaggle dataset, and predictions were made using machine learning algorithms like Nave Bayes, Random Forest, and Decision Tree Classifier. Many models were compared in a predictive analysis, and the Decision Theory classifier was determined to be the most accurate of them, with an accuracy of 97.6%.

**Keywords:** Naive Bayes, Random Forest, Decision Tree Classifier, Machine Learning, Predictive Analysis.

## 1. INTRODUCTION

Heart and blood vessel abnormalities are referred to as cardio vascular illnesses. Those between the ages of 25 and 50 are the most susceptible. Volume 76, Issue 25, published on December 22, 2020, in the journal of the American College of Cardiology, reports that the number of cases of all CVDs has nearly quadrupled from 271 million in 1990 to 523 million in 2019. A sudden disease called a heart stroke is caused by a blockage in the blood arteries that supply the brain. Heart attacks, on the other hand, are brought on by a deficiency in oxygen in the blood that nourishes the heart. Because heart attacks and strokes can happen abruptly, it's crucial to get prompt medical help. A subfield of computer science and artificial intelligence called "machine learning" combines data and algorithms to simulate how people learn, gradually improving the accuracy of the model. To test several machine learning algorithms and methodologies for heart disease prediction, we have created a predictive and comparative model. Cardiovascular disease (CVD) early prediction can help high-risk patients make decisions about lifestyle adjustments, which in turn reduces consequences. With the use of machine learning techniques including Nave Bayes, Random Forest, and Decision Tree Classifier, our suggested model forecasts the risk of heart attacks. Regression analysis is used to properly display each model's outcomes, and a comparison is used to determine which method is the best in terms of accuracy.

## 2. LITERATURE SURVEY

The machine learning approach to find stroke prediction has already been the subject of numerous research papers and publications by scholars. Aayush Shrestha and Tanisha Rakshit [2]. Based on metrics like accuracy, precision, and F1-score, this paper shows the most effective machine learning model utilizing the open-source data set from Kaggle. The outcomes are then tallied and graphically displayed utilizing visualization methods. It splits the data using the cross-validation approach, and the outcomes are as follows: Decision tree: 100% of accuracy, Random Forest: 96.010%, and Naive Bayes: 86.840%. Md. Mahfujur Rahman, Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, M Shamim Al Mamun, and M Shamim Kaiser, [10] To determine the effectiveness of a stroke occurring in a person, eight classifiers were utilized in this study. After the models are evaluated, the results are tabulated with the two attributes "Stroke" and "No stroke," and the optimal algorithm is determined based on accuracy. It uses a heat map to depict the correlationship. For this dataset, the random forest model was discovered to be the most effective. Sonam Nikhar and A.M. Karandikar [6] cited this research on the heart disease prediction system by employing various classifier algorithms. In the paper, decision tree classifier, random forest, and naive bayes machine learning method are compared. The decision tree classifier has superior accuracy than the Naive Bayes classifier, according to the results. Machine learning methods such decision trees, random forests, Naive bayes, SVM, and ensemble models are integrated assessed in a detailed analysis presented by V.V. Ramalingam et al. (9). Their findings show that Naive bayes classifier performance was computationally quick and performed well. While Random Forest and Ensemble Models use numerous decision trees to increase accuracy, they both had good accuracy rates. These results showed that every algorithm has unique characteristics and drawbacks.

## 3. PROPOSED MODEL

This study presents an analysis of several Machine Learning approaches. The algorithms used to accurately identify cardiac disorders including heart attacks and heart strokes include Naive Bayes, Random Forest, and Decision Tree Classifier. The research for this study involves looking through journals, published paper and latest statistics on cardio vascular disease (CVD). The approach is a series of procedures that converts the given data into identifiable data patterns for the knowledge of users. The proposed techniques consist of three stages, the first of which is data collection stage, the second is the stage of significant value extraction and third is the data preparation stage. The data preparation stage deals with the missing values and cleaning of data. The proposed model is then put into practice, and its performance and accuracy are assessed using a variety of performance metrics and visualization techniques, including box plots, which reveal the symmetry, variance, and outliers of the data, histograms, which count the likelihood of each attribute, and heat maps, that are employed to find correlation among two or more attributes. After the evaluation of the algorithmic models their accuracy values are tabulated and are visualized using a Bar chart. They are useful for comparing the values of different categories. The length of the bars indicates the magnitude of the values, and it is easy to compare the lengths of different bars to determine which category has the largest or smallest value. This representation is referred and stated from the published paper: [2].

### A. Model Flowchart

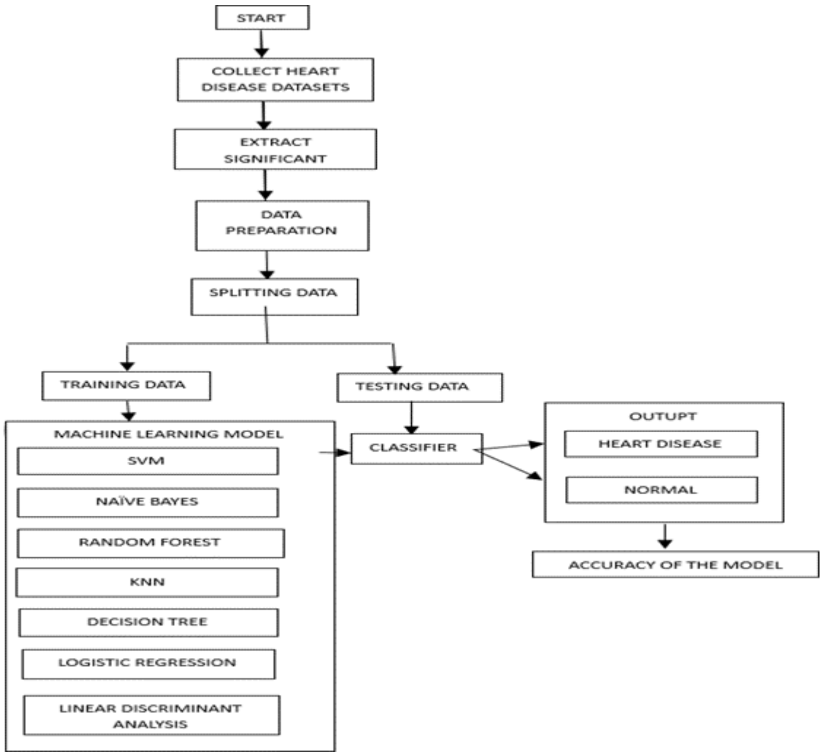
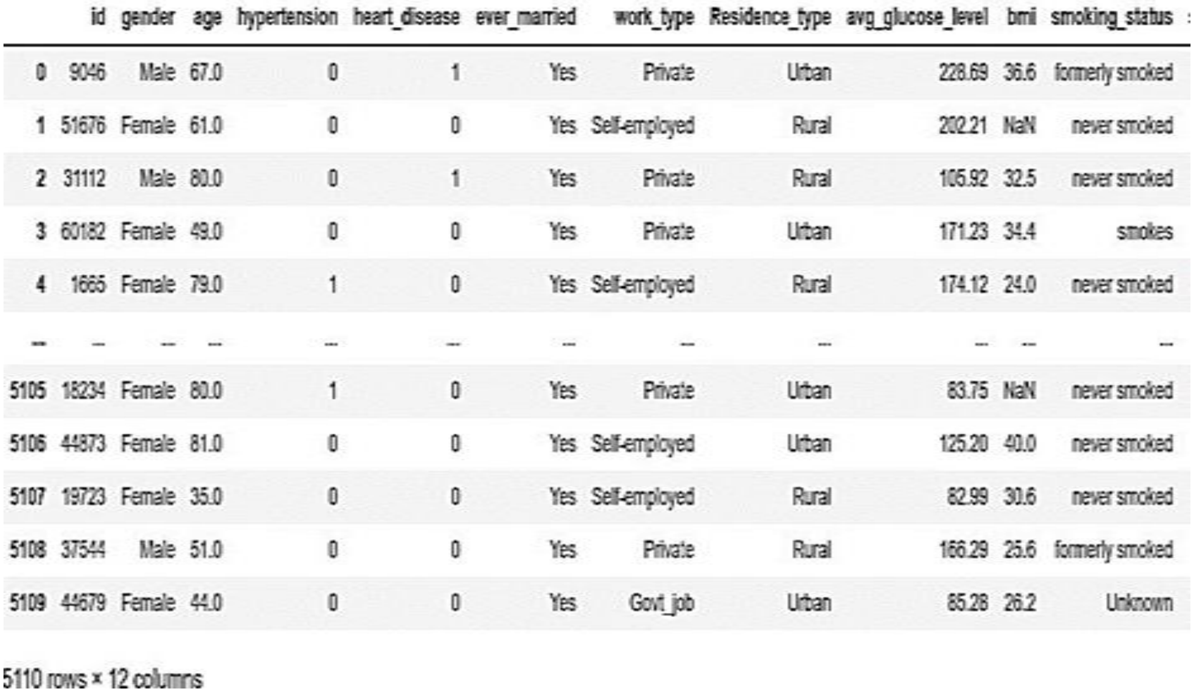


**Figure 1: Flowchart for proposed model**

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 02 | 2025
DOI: 10.5281/zenodo.14842473

### B. Dataset Description

We downloaded the "STROKE PREDICTION DATASET" patient data set from the Kaggle website. There are 12 columns and 5110 rows in this dataset. It has the following 12 characteristics: ID, age, gender, stroke, heart disease, hypertension, ever-married status, work type, residence type, average glucose level, BMI, and smoking status. All of these characteristics are regarded as independent factors, while stroke is regarded as the dependent variable. When employing different machine learning models to perform a predictive analysis on heart stroke, the attribute "id," which represents the patient ID, has no increased relevance. When visualised, the dependent variable is represented on the Y-axis while the independent variables are plotted on the x-axis. Based on the input data used by the machine learning algorithm, this data set is used to determine if a patient is likely to experience a stroke. After comparison on the basis of accuracy and precision, the best machine learning algorithm is found on comparison.



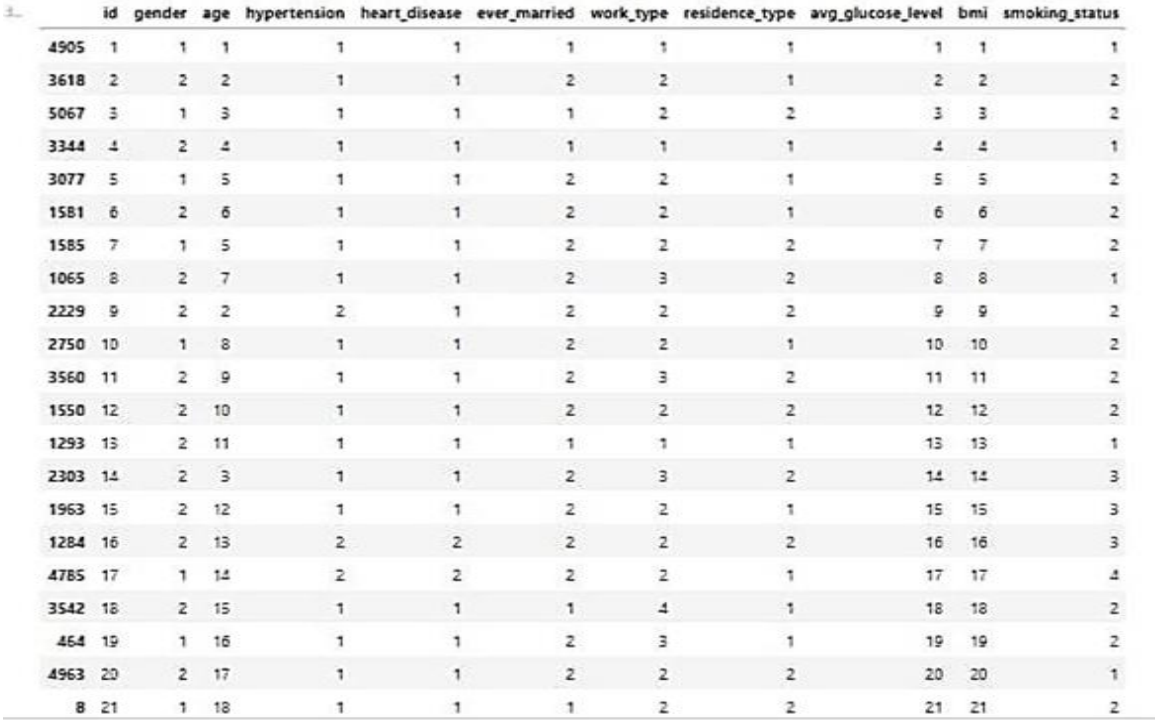
| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 9046 | Male | 67.0 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked |
| 1 | 51676 | Female | 61.0 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | NaN | never smoked |
| 2 | 31112 | Male | 80.0 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked |
| 3 | 60182 | Female | 49.0 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes |
| 4 | 1665 | Female | 79.0 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24.0 | never smoked |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5105 | 18234 | Female | 80.0 | 1 | 0 | Yes | Private | Urban | 83.75 | NaN | never smoked |
| 5106 | 44873 | Female | 81.0 | 0 | 0 | Yes | Self-employed | Urban | 125.20 | 40.0 | never smoked |
| 5107 | 19723 | Female | 35.0 | 0 | 0 | Yes | Self-employed | Rural | 82.99 | 30.6 | never smoked |
| 5108 | 37544 | Male | 51.0 | 0 | 0 | Yes | Private | Rural | 166.29 | 25.6 | formerly smoked |
| 5109 | 44679 | Female | 44.0 | 0 | 0 | Yes | Govt_job | Urban | 85.28 | 26.2 | Unknown |

5110 rows × 12 columns

**Figure 2: Dataset for predicting stroke**

### C. Data Preparation

In the data set we use, the BMI column contains 201 null entries denoted as "NAN" or "None," while the other columns contain 0 null values. A machine learning algorithm's performance and accuracy can suffer if a dataset contains null values. So, it's crucial to handle null values. It is equally crucial to have all the attributes represented numerically in order to do meaningful computations and keep the accuracy of our prediction analysis. So, we treat the null values in the same manner as the category values. The (dropNa) function in the Panda library is used to eliminate the null values or "NAN" under the BMI column. The figure 3. Presented here Represents the input dataset after data preparation process that has only numerical data with no null values or 'NAN'. The shape of dataset after data preparation process is 4909 rows and 12 columns.



| | id | gender | age | hypertension | heart_disease | ever_married | work_type | residence_type | avg_glucose_level | bmi | smoking_status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4905 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3618 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| 5067 | 3 | 1 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 2 |
| 3344 | 4 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 4 | 4 | 1 |
| 3077 | 5 | 1 | 5 | 1 | 1 | 2 | 2 | 1 | 5 | 5 | 2 |
| 1581 | 6 | 2 | 6 | 1 | 1 | 2 | 2 | 1 | 6 | 6 | 2 |
| 1585 | 7 | 1 | 5 | 1 | 1 | 2 | 2 | 2 | 7 | 7 | 2 |
| 1065 | 8 | 2 | 7 | 1 | 1 | 2 | 3 | 2 | 8 | 8 | 1 |
| 2229 | 9 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 9 | 9 | 2 |
| 2750 | 10 | 1 | 8 | 1 | 1 | 2 | 2 | 1 | 10 | 10 | 2 |
| 3560 | 11 | 2 | 9 | 1 | 1 | 2 | 3 | 2 | 11 | 11 | 2 |
| 1550 | 12 | 2 | 10 | 1 | 1 | 2 | 2 | 2 | 12 | 12 | 2 |
| 1293 | 13 | 2 | 11 | 1 | 1 | 1 | 1 | 1 | 13 | 13 | 1 |
| 2303 | 14 | 2 | 3 | 1 | 1 | 2 | 3 | 2 | 14 | 14 | 3 |
| 1963 | 15 | 2 | 12 | 1 | 1 | 2 | 2 | 1 | 15 | 15 | 3 |
| 1284 | 16 | 2 | 13 | 2 | 2 | 2 | 2 | 2 | 16 | 16 | 3 |
| 4785 | 17 | 1 | 14 | 2 | 2 | 2 | 2 | 1 | 17 | 17 | 4 |
| 3542 | 18 | 2 | 15 | 1 | 1 | 1 | 4 | 1 | 18 | 18 | 2 |
| 464 | 19 | 1 | 16 | 1 | 1 | 2 | 3 | 1 | 19 | 19 | 2 |
| 4963 | 20 | 2 | 17 | 1 | 1 | 2 | 2 | 2 | 20 | 20 | 1 |
| 8 | 21 | 1 | 18 | 1 | 1 | 1 | 2 | 2 | 21 | 21 | 2 |

**Figure 3: Dataset after preparation**

## D. Data Visualization

The graphical display of information or data using different visualization techniques, such as graphs, charts, and even animations, is known as data visualization. Complex data is easier to understand thanks to this visual depiction, which also provides insights for data-driven decision-making.

HEAT MAP: They are an effective tool for understanding and displaying complicated data. They aid in the identification of outliers or abnormalities in the data, expose patterns or trends that may not be immediately obvious from a straightforward table or chart, and aid in the determination of the most active or attention- grabbing regions of a dataset.

FEATURE SCORE: By examining the magnitude or position of the feature scores, you can infer the relative importance of different features. Higher scores indicate greater relevance or impact on the target variable.

CLUSTERMAP: It can spot trends, patterns and outliers in the data by examining how the clusters are organized.

RADIAL PLOT: Specific profiles or patterns can be seen in the form made by the interconnected radial lines.

Visualization makes it easier to communicate and interpret data, analyzing past trends and predicting future trends, and to understand relationship between attributes to draw appropriate conclusions.

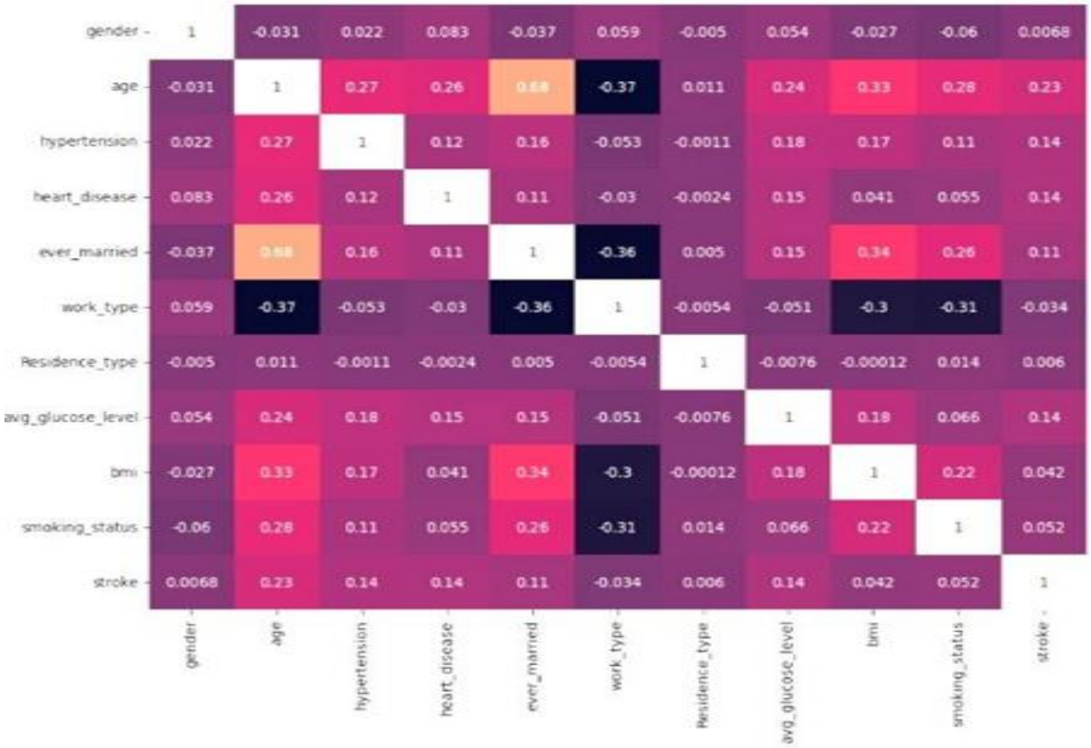


**Figure 4: Heatmap to show trends or patterns in the data**

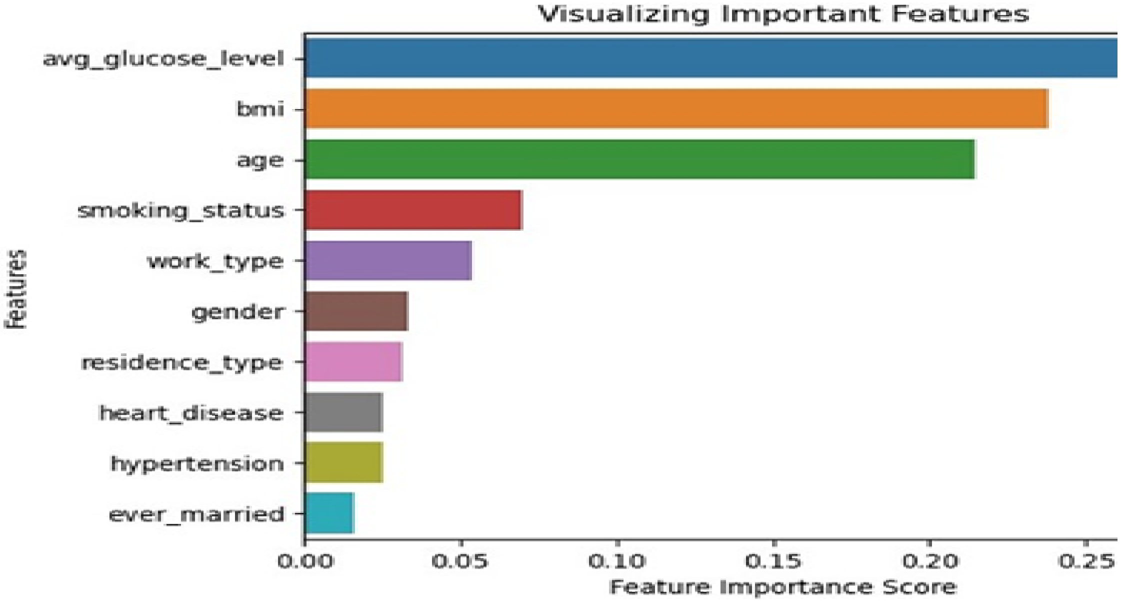


**Figure 5: Feature score relative importance of different features**

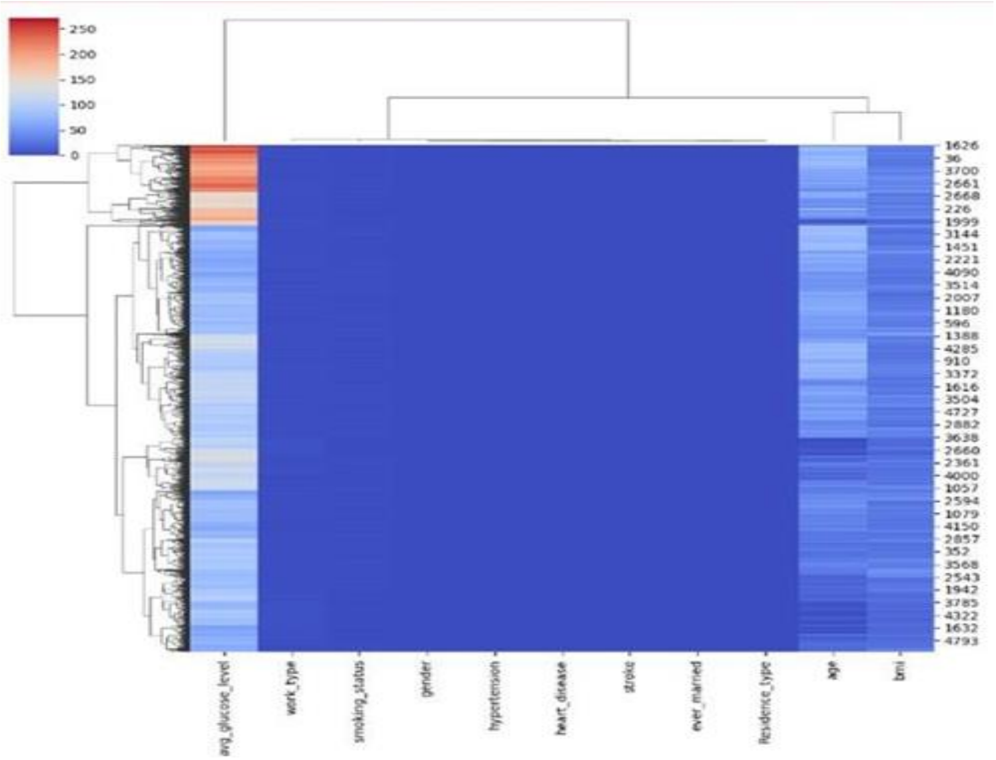


**Figure 6: Cluster map to show trends or patterns in the data**

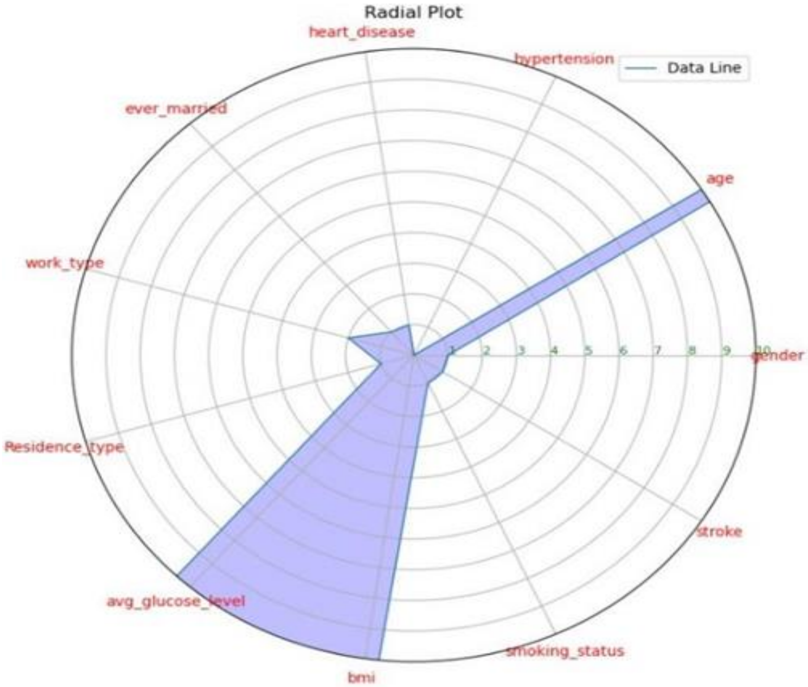


**Figure 7: Radial plot suggests specialization or imbalance in its characteristics**

### E. Data Splitting

The technique of splitting available data into two pieces, typically for cross-validation purposes, is known as data splitting. Both a training model and a testing model are produced. A predictive model is created using a training model, and its performance and accuracy are assessed using a testing model.

The various methods for data splitting includes random subsampling, deterministic methods, for- going data splitting methods and cross validation.

We selected the random subsampling method, which randomly divides the data set into a training set and a testing set with a preferred train size, because our input data is a basic model and smaller dataset. In order to assess a model's effectiveness and accuracy in a comparison analysis, the data must be separated.

### F. Classification

NAÏVE BAYES - The naïve bayes classification algorithm is a supervised learning method that uses the Bayes theorem to resolve classification issues. The probability of each characteristic given the class label may be computed independently, and the probability of the evidence given the hypothesis can then be obtained by multiplying these probabilities together.

It helps develop models with quick prediction skills and is one of the most simple and effective classification techniques. As a probabilistic classification model, it bases its predictions on the likelihood that each object will occur.

According to the Bayes' Theorem, P(A/B), in equation 1, is a representation of the posterior likelihood of an event (A) can be calculated given some prior likelihood of an event (B).

The simple form of the calculation for Bayes theorem is as follows,

$$P(A|B) = (P(B|A)P(A)) / P(B) \qquad (1)$$

RANDOM FOREST – It is a type of ensemble learning method that employs many decision trees to enhance the accuracy and stability of the predictions using a small portion of the training dataset and random selection of attributes.

The objective of a Random Forest classifier is to determine the class of a given data point using its characteristics or qualities. The process creates several decision trees, individually trained on a unique random portion of the data used for training and traits. By distributing the attributes and samples randomly, it is possible to decrease overfitting and boost tree variety.

For each new data point, each tree inside the forest makes a forecast, and the final prediction is chosen by a vote of the majority of the tree predictions in classification tasks or by the average of predictions in regression tasks. It is a versatile and powerful algorithm that can provide high accuracy and reliability for many practical applications.

DECISION TREE CLASSIFIER - A machine learning approach called Decision Tree is utilized for both classification and regression problems. It creates predictions by periodically segmenting the feature set into subsets according to the parameters of the features until each subset has cases that belong to the exact same class or have similar range for the target variable.

In a decision tree, each branch represents a potential value for a feature, and each internal node represents a feature. The Decision tree is built in a top-down manner. The tree is built by repeatedly splitting the data into sections based on the values of the features, up until a stopping condition, such as a root length depth or a minimum number of samples per leaf node, is satisfied.

The model's final output is a forecast predicated upon the majority class or average score of the training cases that reach the leaf node. The outcomes are simpler to read and understand.

They are a straightforward and efficient algorithm that can offer good precision and insights for a variety of real-world applications. Due to its tree- like graph analysis of the dataset, this technique is more accurate than other algorithms.

## 4. RESULTS AND ANALYSIS

After a serious of evaluation of machine learning models for this particular input dataset extracted from an open source such as Kaggle.

For each method independently, the insights and outcomes are recorded in Table 1 below based on some performance measures like accuracy and precision. Many machine learning methods, including the Naive Bayes classifier, decision tree classifier, and random forest models, have been evaluated.

**Table 1: Accuracies Obtained Using Different Algorithms**

| Algorithm | Accuracy |
|---|---|
| NAÏVE BAYES | 80% |
| RANDOM FOREST | 95% |
| DECISION TREE | 97.6% |

The decision tree classifier model, which has the highest accuracy of 97.6%, is followed in accuracy by random forest, which has an accuracy of 95%, and naive bayes classifier, which has an accuracy of 80%.

It is also important to understand that the results may vary depending upon the type of dataset used. For more complex data and large datasets the result could vary to a great extent.

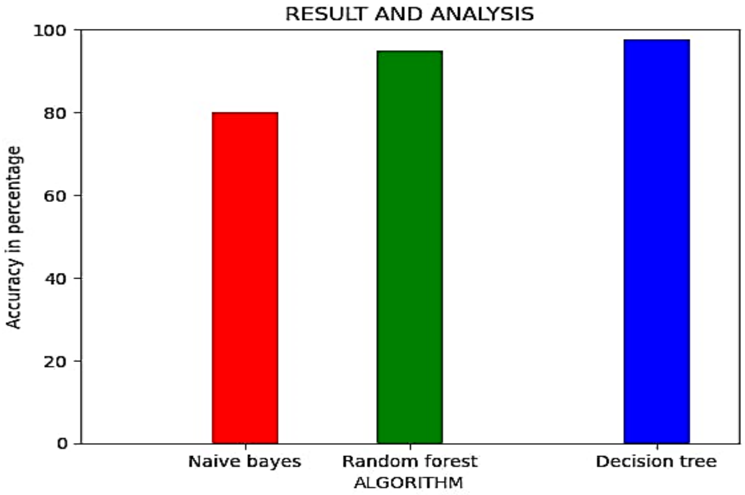


**Figure 8: Bar chart to show Evaluation result of machine learning models**

The results of our evaluation are visually presented to make us understand better. We have used a bar chart to present this. Bar charts are found useful and easy to compare the lengths of different bars and to determine which category has the largest or smallest value.

Thus, our results and findings put forth that decision tree classifier is the most accurate and efficient machine learning algorithm.

## 5. CONCLUSION

As the statistical survey discloses the rate of heart diseases and strokes are increasing rapidly worldwide causing an increase in the mortality rate and it becomes pivotal to find an efficient model with high accuracy rates for any set of input data, that helps to provide proper medical attention.

In our proposed model using random subsampling method decision tree classifier was found to be the most efficient machine learning algorithmic model with effectiveness of 97.6% , and the least effective model was found to be the Naive bayes classifier with effectiveness of 80%. However other accurate models and techniques still exits and some are yet to be discovered. These require an enormous amount of research work and on modeling and evaluation the performance metrics could be enhanced more in the future.

## 6. FUTURE WORK

A big data set is provided as input, and the most accurate outcomes are projected based on comparisons among these models. This work is able to be enhanced by integrating the models together and putting them into an app or a Web-based application for easier user interface and access.

## References

1) Gregory A. Roth, George A. Mensah, Valentin Fuster, " The Global Burden of Cardio Vascular Diseases and Risks", Journal of the American college of Cardiology; Vol. 76 Issue 25, 22- December-2020.

2) Tanisha Rakshit and Aayush Shrestha, "Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT); ISSN: 2278- 0181; published by: htpp://www.ijert.org; Vol. 10 Issue 06, June-2021.

3) Maihul Rajora, "Stroke Prediction Using Machine Learning in a Distributed Environment", published in Springer link: https://link.springer.com/chapter/10.1007/978-3-030-65621-8_15 International Conference on Distributed Computing and Internet Technology; 2021

4) Jaehak Yu, Soon-Hyun Kwon, Sejin park, Kang-Hee Cho and Hansung Lee, "AI-Based Stroke Disease, Institute of Electrical and Electronics Engineering (IEEE), published by: https://ieeexplore.ieee.org; Vol. 10, April-2022.

5) Meng Wang, Xinghua Yao and Yixiang Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients", Institute of Electrical and Electronics Engineering (IEEE), published by: https://ieeexplore.ieee.org; Vol. 9, February-2021.

6) Sonam Nikhar, A.M. Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms" International Journal of Advanced Engineering, Management and Science (IJAEMS), Infogain Publication, Vol-2, Issue-6, June- 2016.

7) I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, Vol. III, T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

8) Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181; published by: htpp://www.ijert.org; Vol. 9 Issue 11, November-2020.

9) V.V. Ramalingam, Ayantan Dandapath and M Karthik Raja," Heart disease prediction using machine learning techniques: a survey", International Journal of Engineering Research and Technology (IJERT), 7 (2.8) (2018) 684-687.

10) Minhaz Uddin Emon , Maria Sultana Keya , Tamara Islam Meghla , Md. Mahfujur Rahman , M Shamim Al Mamun , and M Shamim Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction", Institute of Electrical and Electronics Engineering (IEEE) , published by: https://ieeexplore.ieee.org ; ISBN: 978-1-7281-6387- 1, 2020.

11) Riddhi Kasabe, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org; ISSN: 2278-0181; Vol. 9 Issue 08, August-2020.

12) Apurb Rajdhan, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181; http://www.ijert.org; Vol. 9 Issue 04, April-2020.

13) N. Komal Kumar, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", IEEE Xplore, Published in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS).

14) Kaggle dataset link: https://www.kaggle.com/datasets/fedesoriano/stroke- prediction-dataset